Supplementary Material for "Clustering in the Presence of Background Noise"

Shai Ben-David SHAI@CS.UWATERLOO.CA
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1 CANADA

Nika Haghtalab NIKA@CMU.EDU
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Proof of Theorems and Lemmas

Lemma 1. *Given a (k, g)-centroid algorithm $\mathcal{A}$ and parameter δ, let $\mathcal{A}' = RT_\delta(\mathcal{A})$. For any $\mathcal{I} \subseteq \mathcal{X}$, such that for all $y \in \mathcal{I}$, $d(y, \mu'(y)) \leq \delta$, $\mathcal{I}$ is $g(\delta)|\mathcal{X} \setminus \mathcal{I}|$-cost-robust to $\mathcal{X} \setminus \mathcal{I}$ with respect to $RT_\delta(\mathcal{A})$ for the Λ_d^g (cost) function.*

Proof. If $d(y, \mu'(y)) \leq \delta$, then y is not in the noise cluster. Therefore, there is a cluster $C^* \in \mathcal{C}$, such that $C^* \subseteq \mathcal{X} \setminus \mathcal{I}$, and for any $C' \neq C^*$,

$$\Lambda_d^g(C') \leq \Lambda_d^g(C' \cap \mathcal{I}) + |C' \setminus \mathcal{I}| \cdot g(\delta)$$

Therefore, $\mathcal{I}$ is $g(\delta)|\mathcal{X} \setminus \mathcal{I}|$-cost-robust to $|\mathcal{X} \setminus \mathcal{I}|$ with respect to $RT_\delta(\mathcal{A})$ for the Λ_d^g cost function. □

Lemma 3. *Let $\mathcal{A}$ be the (k, g)-centroid algorithm. For any $\mathcal{I}$, if it can be covered with a (ρ_1, ρ_2)-balanced set of k balls, called $\mathcal{B}$, where each ball has radius r and the centers of two different balls are at least $\nu > 4r + 2g^{-1}(\frac{\rho_1+\rho_2}{\rho_1} g(r))$ apart, then $\mathcal{A}(\mathcal{I}) = \mathcal{B}$.*

Proof. Let $\mathcal{B} = \{B_1, \ldots, B_k\}$ and for $i \in [k]$ let b_i represent the center of B_i and D_i represent a ball of radius $\frac{\nu}{2} - r$ centered at b_i. Let $\mathcal{A}(\mathcal{X}) = \mathcal{C}$ with centers $\mu_1, \ldots, \mu_k$. Let $\mathcal{D}_1 = \{D_i |\, D_i \text{ does not cover any } \mu_j\}$ and $\mathcal{D}_2 = \{D_i |\, D_i \text{ covers more than one } \mu_j\}$. Since $D_1, \ldots, D_k$ are pairwise disjoint, $|\mathcal{D}_1| \geq |\mathcal{D}_2|$. Assume in search of a contradiction that $\mathcal{D}_1 \neq \emptyset$. For any $D_i \in \mathcal{D}_1$, for all $y \in D_i$, $d(y, \mu(y)) \geq \frac{\nu}{2} - 2r$. Consider the following set of $\mu_1'', \ldots, \mu_k''$: If D_j includes exactly one center, μ_i, then let $\mu_j'' = \mu_i$, otherwise $\mu_j'' = b_j$.

$$\begin{aligned}
\Lambda_{\mathcal{I},d}^g(\mu_1'', \ldots, \mu_k'') &\leq \Lambda_{\mathcal{I},d}^g(\mu_1, \ldots, \mu_k) \\
&\quad + \sum_{D_i \in \mathcal{D}_1} \sum_{y \in B_i} [g(d(y, \mu''(y))) - g(d(y, \mu(y)))] \\
&\quad + \sum_{D_i \in \mathcal{D}_2} \sum_{y \in B_i} [g(d(y, \mu''(y))) - g(d(y, \mu(y)))]
\end{aligned}$$

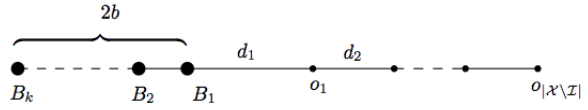

Figure 1. Structure of a data set that is not robust w.r.t $RI_p(\mathcal{A})$.

$$\begin{aligned}
&\leq \Lambda_{\mathcal{I},d}^g(\mu_1, \ldots, \mu_k) + \sum_{D_i \in \mathcal{D}_1} |B_i| \left(g(r) - g(\frac{\nu}{2} - 2r)\right) \\
&\quad + \sum_{D_i \in \mathcal{D}_2} |B_i| g(r) \\
&\leq \Lambda_{\mathcal{I},d}^g(\mu_1, \ldots, \mu_k) + \rho_1 |\mathcal{D}_1||\mathcal{I}| \left(g(r) - g(\frac{\nu}{2} - 2r)\right) \\
&\quad + \rho_2 |\mathcal{D}_2||\mathcal{I}| g(r) \\
&\leq \Lambda_{\mathcal{I},d}^g(\mu_1, \ldots, \mu_k) \\
&\quad + |\mathcal{D}_1||\mathcal{I}| \left((\rho_1 + \rho_2) g(r) - \rho_1 g(\frac{\nu}{2} - 2r)\right) \\
&< \Lambda_{\mathcal{I},d}^g(\mu_1, \ldots, \mu_k)
\end{aligned}$$

This forms a contradiction, so without loss of generality every D_i covers a center μ_i. For $i \neq j$ and for all $y \in B_i$, $d(y, \mu_i) \leq \frac{\nu}{2} < d(y, \mu_j)$. Therefore, $\mathcal{A}(\mathcal{I}) = \mathcal{B}$. □

Theorem 6. *Let $\mathcal{A}$ be the k-means algorithm. For any $r, \lambda > 0$, there exists $\mathcal{X}$ and $\mathcal{I} \subseteq \mathcal{X}$, such that $\mathrm{rad}(\mathcal{I}) \leq r$, $\mathcal{I}$ can be covered with k balls of arbitrarily small radii, and $\mathcal{X}$ has signal-to-noise ratio of $\frac{|\mathcal{I}|}{|\mathcal{X} \setminus \mathcal{I}|} \geq \lambda$. But, for any $p < |\mathcal{X} \setminus \mathcal{I}|$, $\mathcal{I}$ is not $(1 - \frac{1}{k})$-robust with respect to $RI_p(\mathcal{A})$.*

Proof. We repeat the construction from Theorem 2 and Figure 1 with $\alpha = 2r$. Let $d_1 = 4r(\frac{\lambda}{\lambda+1}|\mathcal{X}| + 1)$ and $d_2 = 2(d_1 + 2r) + 1$. For $i \in [k]$, let B_i denote a set with radius 0, such that $|B_i| \geq \frac{\lambda}{k(\lambda+1)}|\mathcal{X}|$. Let $B_1, \ldots, B_k$ be evenly placed on a line of length $2r$. For, $i \in \left[\left\lfloor \frac{|\mathcal{X}|}{\lambda+1} \right\rfloor\right]$, let o_i be a point on the line that connects $B_1, \ldots, B_k$, such that $d(o_1, B_1) = d_1$ and $d(o_i, o_{i+1}) = d_2$ (see Figure 1). Let $\mathcal{I} = \bigcup_{i \in [k]} B_i$ and $\mathcal{X} = \mathcal{I} \cup \{o_1, \ldots, o_{|\mathcal{X}|/(\lambda+1)}\}$. Note that $\mathcal{X}$ and $\mathcal{I}$ are chosen such that $\frac{|\mathcal{I}|}{|\mathcal{X} \setminus \mathcal{I}|} \geq \lambda$.

Similar to Theorem 2, we have that for all $y \in \mathcal{I}$, $d(y, \mu(y)) > 2r$. Since the clusters in any centroid-based clustering are convex, the center of the cluster containing any B_i is to the right of $\mathcal{I}$ (see Figure 1). Therefore, $B_1, \ldots, B_k$ are all in one cluster of $RI_p(\mathcal{A})(\mathcal{X})$. Each B_i forms a unique cluster in $\mathcal{A}(\mathcal{I})$. Therefore,

$$\begin{aligned}
\Delta(\mathcal{A}(\mathcal{I}), \mathcal{A}'(\mathcal{X})|\mathcal{I}) &\geq 1 - \frac{\sum_{i\in[k]} \binom{|B_i|}{2}}{\binom{|\mathcal{I}|}{2}} \\
&\geq 1 - \frac{k\binom{\frac{|\mathcal{I}|}{k}}{2}}{\binom{|\mathcal{I}|}{2}} \\
&\geq 1 - \frac{1}{k}
\end{aligned}$$

$\mathcal{I}$ is not $(1-\frac{1}{k})$-robust to $\mathcal{X}\backslash\mathcal{I}$ with respect to $RI_p(\mathcal{A})$. $\square$